研报情感分析及分类

赵子衡（量化实习）

目录

CONCENTS

ONE

数据预处理

保留中文字符并对文档进行分词处理

数据预处理

保留中文字符

01

考虑到原始数据中存在较多噪音数据，且后续情感分析仅以中文单词为基础，因此文档中仅**保留中文字符**，去除噪音文本数据。

中文分词

02

设置停用词并补充自定义词典，去除无意义的中文字符并避免专业术语被错误划分；使用jieba中文分词库，将单篇文档划分为词列表。

TWO

情感得分模型

情感得分模型以中文词语为单位，依据SO-PMI算法得出词语情感倾向得分，基于TF-IDF算法得出词语重要度得分，然后基于重要度得分对情感度得分进行加权得到文档的情感得分。

情感得分模型

01

SO-PMI算法

基于文档词汇列表与**种子词汇表**（正/负面词典），分别计算未知词汇与正/负面词汇的**PMI（点互信息），**求和做差即得到**SO-PMI**，作为**情感倾向得分**。

02

TF-IDF算法

TF-IDF是一种针对**关键词**的统计分析方法，用于评估一个词对一个文件集或者一个语料库的重要程度。**一个词的重要程度跟它在文章中出现的次数成正比，跟它在语料库出现的次数成反比。**我们以TF-IDF得分作为**重要度得分。**

03

情感得分

基于单词**重要度(TF-IDF)得分**对**情感度(SO-PMI)得分**进行加权得到文档的情感得分。然后利用**sigmoid函数**将文档原始情感得分映射到0-10区间内。情感得分>5为积极，<5则为消极。

SO-PMI

Seed dictionary

Bian, Shibo & Jia, Dekui & Li, Feng & Yan, Zhipeng. (2019). A New Chinese Financial Sentiment Dictionary for Textual Analysis in Accounting and Finance. SSRN Electronic Journal. 10.2139/ssrn.3446388.

01

Pointwise Mutual Information

PMI（点互信息）用来衡量两个事物（词汇）之间的相关性。

$$log_2 \frac{P(A,B)}{P(A)P(B)}$$

P(A):词汇A出现的文章的概率
P(B):词汇B出现的文章的概率
P(A, B):词汇AB一起出现在文章中的概率

02

SO-PMI

对于词汇A，计算其情感倾向得分：

$$\sum_{pw \in POS} (PMI(A,pw)) - \sum_{nw \in NEG} (PMI(A,nw))$$

pw： positive word (积极词汇)
nw： negative word (消极词汇)

03

Seed dictionary

A New Chinese Financial Sentiment Dictionary for Textual Analysis in Accounting and Finance

Step 1

整 合 **HOWNET, DLUTSD and NTUSD**三个情感词典，并移除重复词汇.

Step 2

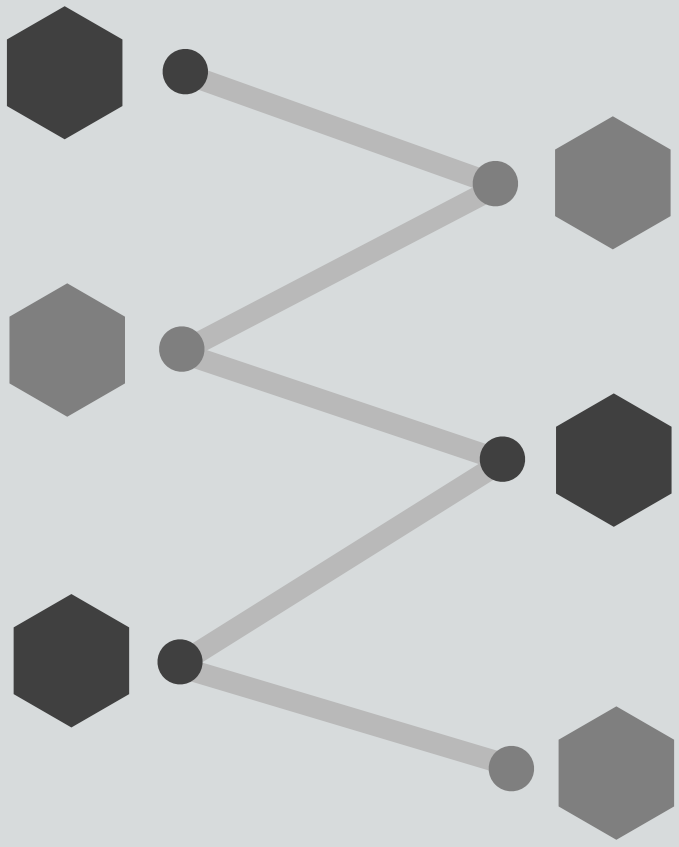
收集**1,411份线上路演纪要，7,138 电话会议纪要，2,043 IPO招股书，29,737 份年报.**使用Jieba包对以上语料进行分词。

Step 3

去除基础语料中不包含的情感词汇

Step 4

人为增加100个金融领域常用的积极词汇与100个消极词汇100作为CFSD 0.1.



Step 5

人为去除词典中与金融领域无关的词汇及其相似词汇。

Step 6

最终得到1,488个负面词汇与1,107个正面词汇。

Seed dictionary

Appendix, Part I: CFSD Positive Word List

English	Chinese
A hundred flowers bloom	百花齐放
A person of outstanding talent	俊杰
Abide by	信守
Abundance	丰盈
Abundant	充沛
Academician	院士
Accepted	公认
Accessible from all directions	四通八达
Acclaim	赞誉
Accuracy	精确性
Accurate	精确
Accurate	准确
Accurate	准确无误
Achieve	达到
Achieve	实现
Achieve something with glory	荣登
Achieve success	建功立业
Achievement	成就
Act with united strength	通力合作
Active	活跃
Actively	踊跃
Adapt	适应
Adapt to local conditions	因地制宜
Adequate	充足
Adhere to	坚持
Appropriate	适当
Appropriate	妥当
As always	一如既往
As one wishes	如意
Ascend	跻身
Assist	协助
At the top of the list	名列前茅
Attract	吸引
Attractive	吸引力
Authentic	正品
Authoritative	权威性
Authority	权威
Autonomous	自主
Autonomy	自主性
Avoid	避免
Backbone	中坚
Balanced	均衡
Be able to accomplish great deeds	大有作为
Be bold and enterprising	勇于进取
Be chosen	中选
Be enthusiastic and press on	奋发有为
Be fair in meting out rewards or punishments	奖罚分明
Be in the full vigor of life	年富力强
Be in the same boat	同舟共济
Be prepared for	
Blessed land	福地
Bold	大胆
Boldness of vision	气魄
Bole	伯乐
Bona fide	善意
Boom	景气
Boom in production and sales	产销两旺
Booming	红火
Booming	蒸蒸日上
Boutique	精品
Boutiqueization	精品化
Brainstorming	集思广益
Brand new	全新
Brand new	崭新
Brave	勇敢
Brave	勇于
Breakthrough	突破
Breakthrough	突破性
Bright	璀璨
Bright	光明
Bright	明亮
Bright and beautiful	亮丽
Brilliant	灿烂
Brilliant	光辉
Brilliant	辉煌
Bring together	凝聚
Brisk buying and selling	购销两旺
Cautious	谨慎
Celebrity	名流
Certainty	确定性
Certificate of merit	奖状
Champion	冠军
Change the appearance of	改观
Change with each passing day	日新月异
Charitable	慈善
Charm	魅力
Charter	特许
Cheap, but good	价廉物美
Cheap, but good	物美价廉
Cherish	珍惜
Chrish	珍视
Civilization	文明
Classic	经典
Clean	洁净
Clean	清洁
Clean	清正
Clean and hygienic	清洁卫生
Clean government	廉政
Clear	明朗
Clear	明确
Clear	明晰
Clear	清楚
Clear	清晰
Clear thinking	思路清晰
Confident	有把握
Conquer	攻克
conscientious	认真
conscientiously	认认真真
Considerable	可观
Considerable	长足
Consideration	兼顾
Consistent	一致
Consolidate	巩固
Constructive	建设性
Contribute	添砖加瓦
Contribution	建树
Contribution	贡献
Convenience	方便
Convenient	便捷
Convenient	便利
Convinced	确信
Cooperate with absolute sincerity	精诚团结
Cooperation	合作
Cooperation	协作
Cooperation	协力
Coordination	协调
Core	核心
Correct	正确
Correction	矫正
Correctness	正确性
Courage	魄力
Courageous	奋勇

Seed dictionary

Appendix, Part II: CFSD Negative Word List

English	Chinese
A flash in the pan	昙花一现
A rush for quick results	急于求成
A sudden turn for the worse	急转直下
Abandon	抛弃
Abandon	废弃
Abnormal	异常
Abort	中止
Abruptly	陡然
Absence	缺席
Absurd	荒唐
Abuse	滥用
Abuse of power	滥用职权
Abuse one's power to seek personal gain	以权谋私
Accept bribes	受贿
Accident	事故
Accident	意外
Accident	肇事
Accident	意外事故
Accident	意外事件
Accountability	追责
Accusation	指责
Accuse	控告
Adulteration	掺假
Advance rashly	冒进
Adverse consequences	不良后果
Arrogant	自大
Ashamed	惭愧
Ashamed	汗颜
Ashamed	羞愧
Ashamed	有愧
Asking for exorbitant price	漫天要价
Attached	依附
Attack	发作
Attack	攻击
Attack	袭击
Attempt	企图
Autarchy	一言堂
Avalanche	崩落
Avoid	避开
Avoid	回避
Avoid answering	避而不答
Avoid talking	避而不谈
Awkward	尴尬
Backlog	积压
Backlog of supplies	积压物资
Bad	不良
Bad	恶劣
Bad	糟糕
Bad	差劲
Bad feelings	过节
Bad news	负面新闻
Bad thing	坏事
Be perfunctory	敷衍了事
Be punished	受罚
Bear	承受
Beat	击败
Beat down the price	杀价
Become rigid	僵化
Behind schedule	误期
Believe what one hears	听信
Betray	出卖
Bias	偏心
Biased	偏向
Big talk	大话
Big-pot Distribution System (extreme equalitarianism)	大锅饭
Bitter	惨痛
Black box	暗箱
Black curtain	黑幕
Blame fate and other people	怨天尤人
Blend	掺和
Blind	盲目
Blind development	盲目发展
Blind investment	盲目投资
Blind optimism	盲目乐观
Blind production	盲目生产
Blindly	一味
Blindness	盲目性
Bubble	泡沫化
Bubble	泡沫
Bulky	笨重
Bulletin criticism	通报批评
Burden	包袱
Burden	负担
Bureaucracy	官僚主义
Bureaucratic	打官腔
Bureaucratic tone	官话
Bureaucratic tone	官腔
Burn	烧毁
Burst	突发
Business losses	企业亏损
Buy winners and sell losers blindly	追涨杀跌
Calamity	灭顶之灾
Calculating	算计
Camouflage	伪装
Can not	没法
Can not	无法
Can not be ignored	不容忽视
Can not get it	得不到
Can not recover	收不回
Can not solve	解决不了
Cancel	取消
Cancellation	解约
Cannot help but	无奈
Capricious	任性

TF-IDF

Term Frequency

$$TF_W = \frac{\text{在某一文档词条}w\text{出现的次数}}{\text{该文档中所有的词条数目}}$$

词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。

01

Inverse Document Frequency

$$IDF_W = log(\frac{\text{语料库中的文档总数目}}{\text{包含词条w的文档数目}+1})$$

逆向文件频率 (inverse document frequency)
IDF的主要思想是：如果包含词条t的文档越少，**IDF越大，则说明词条具有很好的类别区分能力。**

02

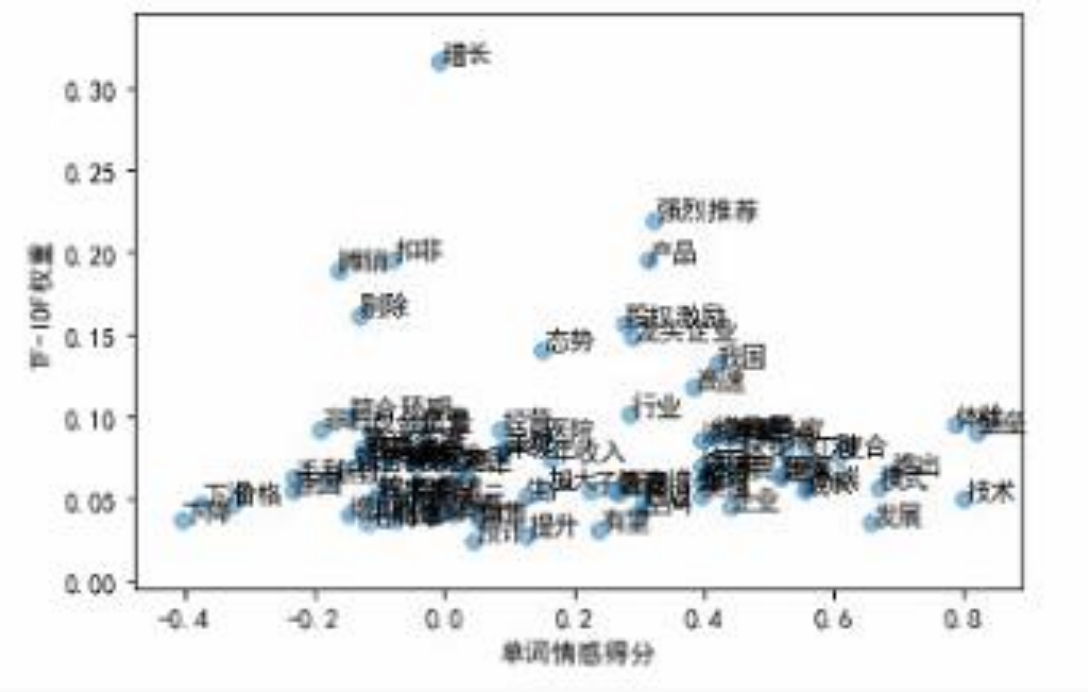
TF-IDF

$$TF-IDF_W = TF_W * IDF_W$$

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

03

情感得分模型



研报情感得分：7.60

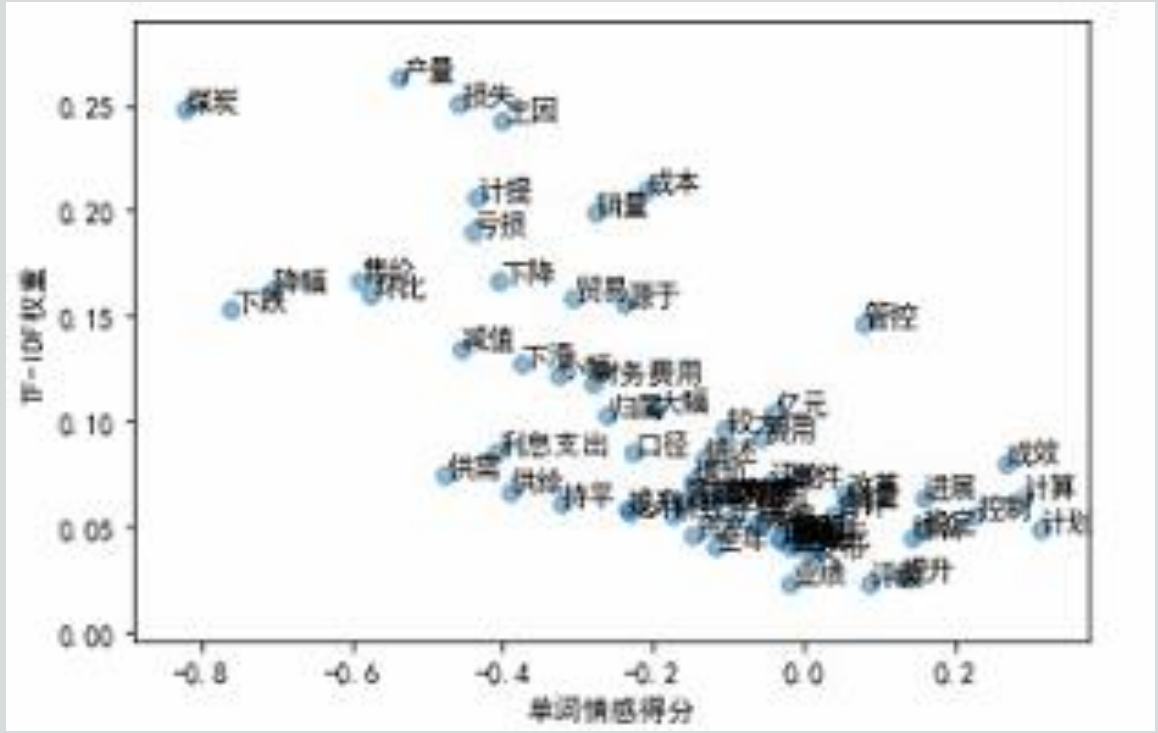
事件：公司发布2018年第一季度财务报告，营业收入同比增长33.25%，归母净：利润同比增长32.08%，扣非后归母净利润同比增长20.03%，剔除Q1摊销**股权激励**费用310万，扣非后归母净利润同比增长32.33%，完全符合预期。维持"**强烈推荐**-A"评级。

Q1业绩继续高增长。18年Q1营业收入0.77亿元，同比增长33.25%，归母净利润0.36亿元，同比增长32.08%，经营活动产生的现金流量净额0.38亿元，同比增长38.08%，业绩继续保持高增长，主要原因是一季度包含了中小学的寒假，而寒假又是角膜接触镜佩戴的"高峰期"。此外公司扣非后归母净利润0.31亿元，同比增长20.03%，剔除股权激励摊销的310万，公司扣非后归母净利润同比增长32.33%。毛利率同比**下降**2pct，判断是低毛利的护理液产品占比有所提升；销售费用率同比**下滑**1.27pct，公司Q1销售费用同比增长22.34%，管理费同比增长107.84%，较大的增幅，主要是由于17年公司新增加12家子公司所致。

龙头企业，充分享受行业高成长**红利**：公司是我国大陆地区目前唯一获得国家食药监总局颁发的角膜塑形镜产品注册证的生产企业。由于角膜塑形镜**技术**和安全要求较高，产品准入许可制度较为严格，行业**壁垒**高；公司未来3-5年龙头企业地位稳固，由于我国近视人群众多，但硬性角膜渗透率较低，与欧美国家相比仍有很大的空间，2017年公司的产品已经进入700多家医院验配点，累计验配超过60万例。18年公司将继续加大终端渠道的整合；同时自产产品镜特舒冲洗液也呈现**高速增长**态势，18年有望继续维持。公司2012-2017年收入年复合增长率35.5%，呈高速发展态势。公司18年元月推出的全新产品DreamVision，以其较高的价格和更舒适的佩戴体验，将为公司吸引更多的**高端**消费人群也为公司贡献更多利润。

维持"**强烈推荐**-A"评级。预计公司2018-2020年归母净利润增速分别为35%/28%/28%，对应EPS分别为1.63/2.08/2.67元，看好公司行业龙头属性，和未来的眼视光诊疗一体化的经营模式，继续维持"**强烈推荐**-A"评级。<!- 欧普康视（300595.SZ）：Q1经营性扣非净利润快速增长32%符合预期- >

情感得分模型

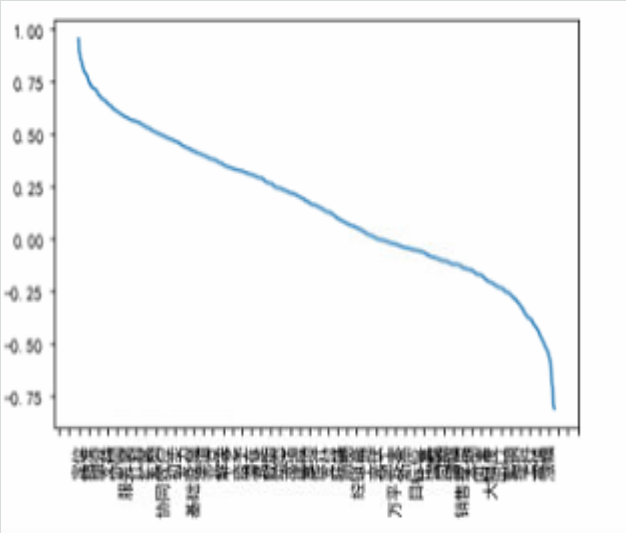


研报情感得分：1.43

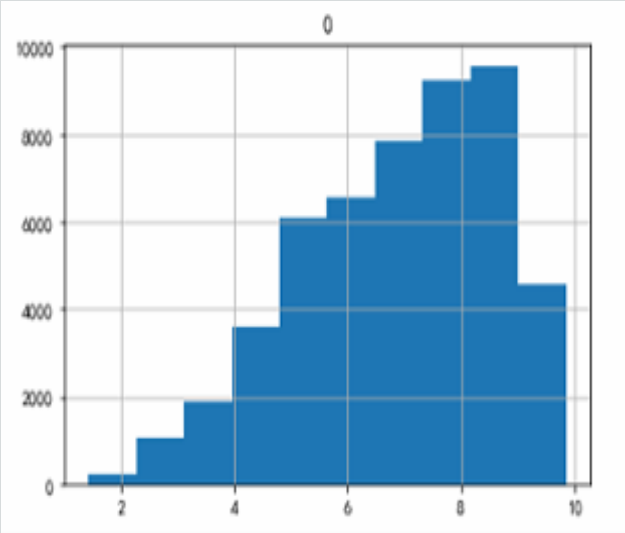
事件描述：公司发布2015年报和2016年1季报，分别实现EPS-1.20元和-0.03元。

事件评论：2015年原煤产量同比增长，销量下滑主因贸易量下滑。2015年公司原煤产量1953万吨，同比增长14.88%，其中自产煤销量1774万吨。分矿井看，经坊矿和大平矿产量同比分别下降4.47%、15.03%，鹿台山和长春兴矿投产，贡献主要增量。15年煤炭总销量6644.39万吨，同比下降38.20%，销量同比降幅较大主因贸易煤量下滑。2015年煤炭发运量2766.14万吨，其中铁运量1168.14万吨，汽运量1598万吨。2016年公司计划原煤产量2000万吨，同比小幅提升。成本管控较好缓冲弱势煤价冲击。2015年煤炭市场不景气，供需宽松下煤价大幅下跌，公司也难以独善其身，吨煤综合平均售价同比下降了83.61元。但成本端管控取得一定成效，吨原煤完全成本133元，同比下降36%。以原煤产量口径计算，2015年吨煤售价194元，同比下跌25.66%，吨煤成本同比下降38.09%，吨煤毛利86.34元，同比基本持平。2015年煤炭开采毛利率同比上升11.15个百分点，弱势行情下成本控制重要性不言而喻。2015年贸易煤毛利率小幅降至1.38%。　　费用整体稳定，管理费用降幅较大。2015年公司期间费用合计32.19亿元，同比上升1.35%，其中销售费用同比上升12.7%，主要是运输港杂费增加所致，财务费用同比上升源于利息支出上升。　　4季度业绩亏损源于计提减值损失，1季度环比减亏。2015年4季度公司归属净利润亏损22.91亿元，主因财务费用环比大幅增加及计提资产减值损失8.31亿元（主要是坏账损失，全年计提坏账损失7.37亿元）。1季度公司归属净利润亏损0.50亿元，环比大幅减亏。关注诉讼及山西供给侧改革进展。预测公司2016-2018年EPS分别为0.09、0.10、0.11元，维持"增持"评级。

情感得分模型



词汇情感得分分布



研报情感得分分布

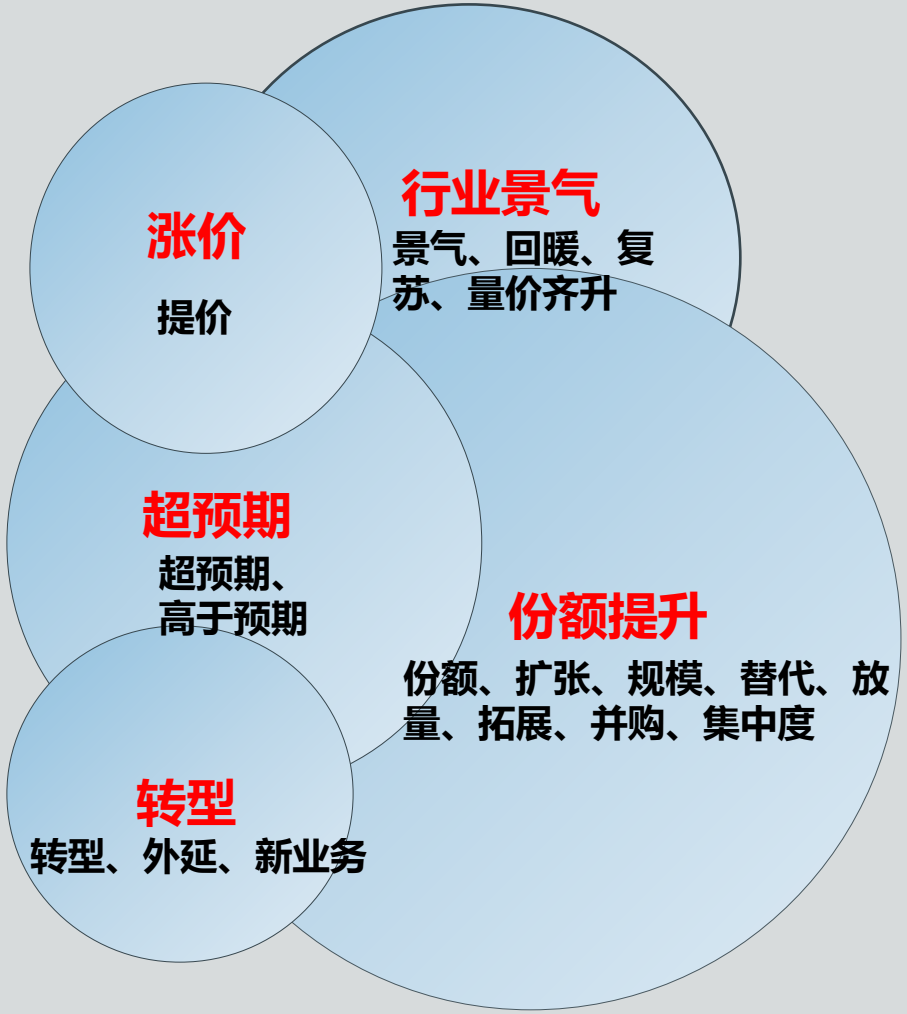
由于国内没有完善的做空机制，看空或者消极情绪的研报极少，因此情感得分分布明显右偏，得分主要集中于6-9分之间。后续我们可以通过对消极词汇进行惩罚，改善情感得分分布的右偏状况。

THREE

研报分类模型

利用关键词检索的形式确定研报分类以及基于监督学习算法预测

研报分类模型--关键词索引



从价量两方面考虑：

涨价->需求侧（行业景气度回升）

供给侧（行业集中度提高）

量增->需求侧（行业景气度回升）

供给侧（规模/产能/份额扩张、并购及行业集中度提高等）

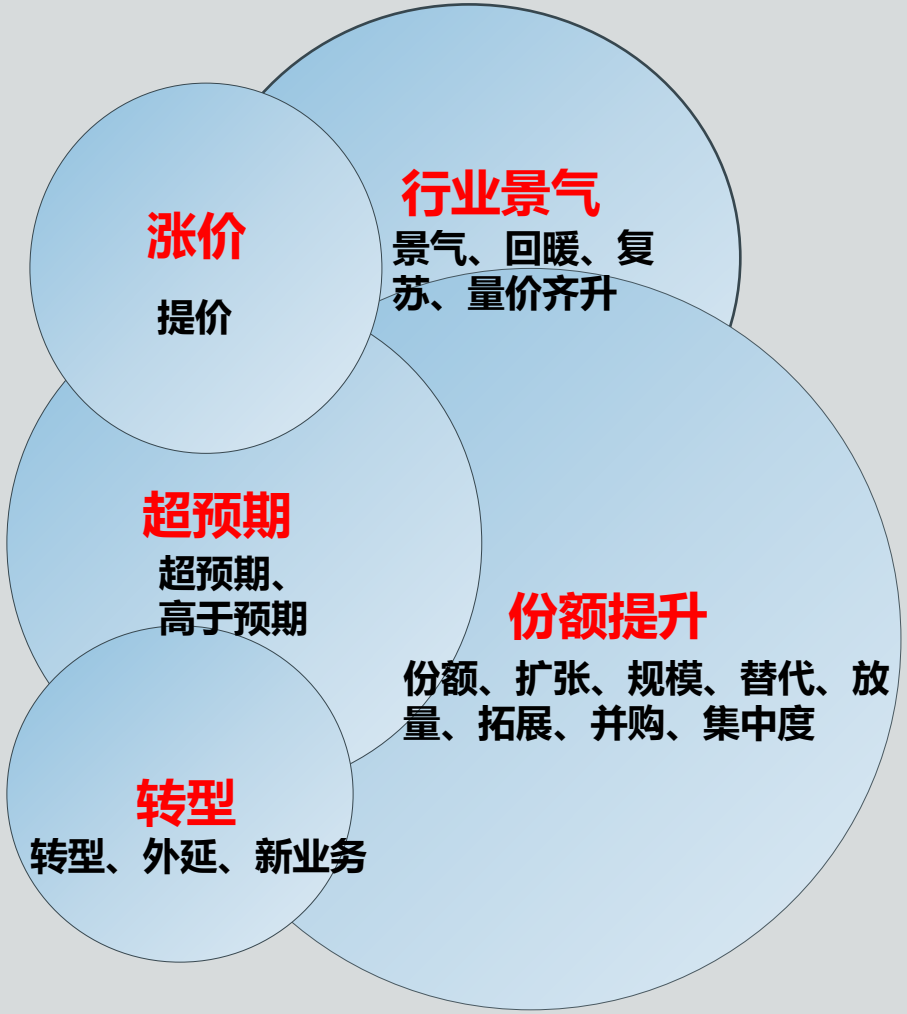
行业景气->需求侧（需求旺盛）

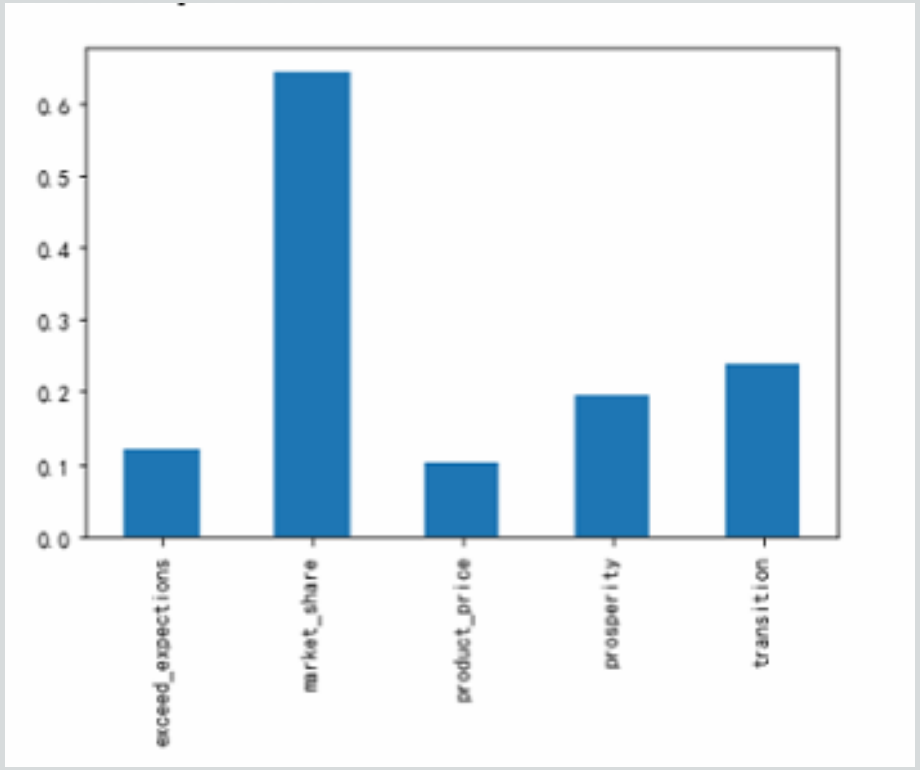
->供给侧（行业集中度提高，竞争格局改善）

转型-> 战略转型/外延并购

超预期->业绩超预期

研报分类模型

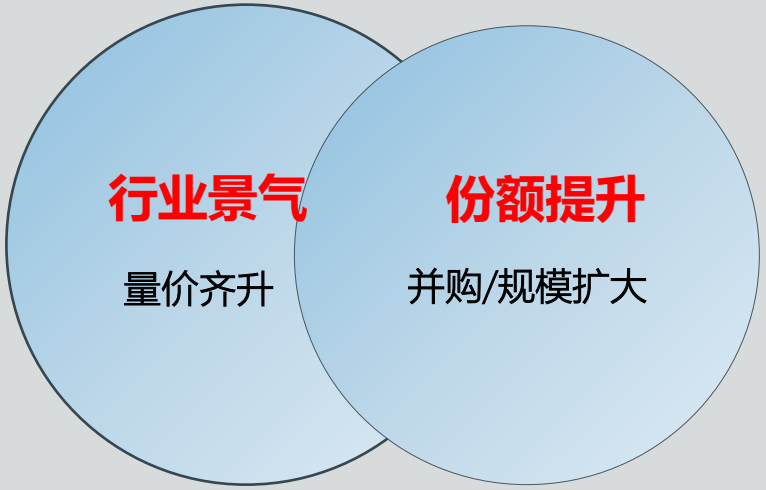


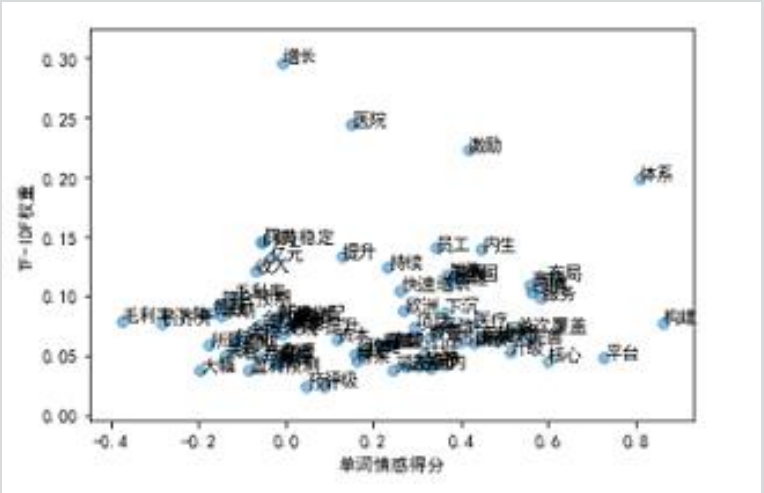


文本类别分布

研报分类模型--案例分析

好赛道+龙头公司





综合情感得分：7.97
行业景气得分：3.99
份额提升得分：3.99

报告摘要：爱尔眼科发布2017年年报：实现营业收入</AI>59.63亿元，同比增长49.06%；归母净利润</AI>7.43亿元、归母扣非净利润7.76亿元，分别比同期增长33.31%、41.87%。同时发布2017年利润分配预案：每10股派发现金红利3元（含税），以资本公积金每10股转增5股。

业绩符合预期，维持高速增长。公司核心医疗服务项目内生增长强劲，其中屈光项目收入19.31亿元，同比增长69.23%，毛利率52.81%，下降1.54pct。屈光手术的增长一方面是境内各医院手术量快速增长的同时全飞秒、ICL等高端手术占比进一步大幅提高，形成**量价齐升**；另一方面本期**并购**欧洲ClínicaBaviera.S.A经营**规模扩大**所致。白内障项目收入14.17亿元，同比增长44.43%，毛利率38.21%，增长0.93pct。主要是受白内障市场的快速增长以及高端多焦晶体，全飞秒术式的应用带来收入的大幅提升；视光服务项目收入11.72亿元，同比增长34.36%，毛利率53.66%，提升1.39pct；眼前、后段手术分别同比增长29.73%和44.16%，毛利率保持稳定。整体来看，公司内生稳健，毛利率稳定，费用可控制得宜，保持稳定高速增长。

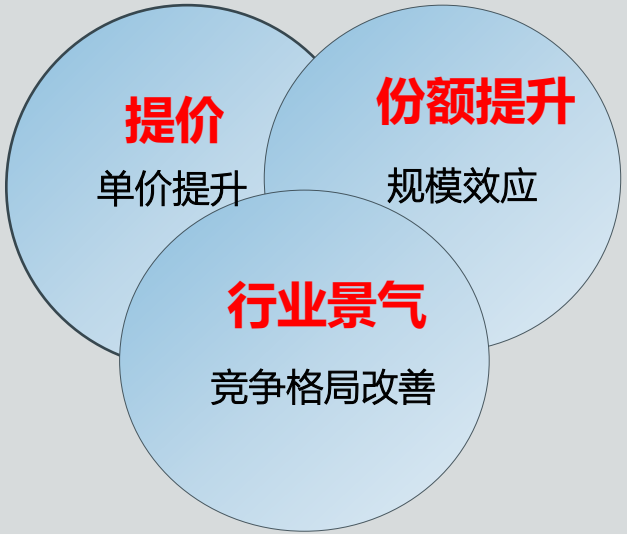
眼科市场空间巨大，全国分级连锁布局保障服务人次不断提升。公司继续加快全国分级连锁网络布局，通过新建或**并购**方式加快地级、县级医院的网点纵向布局，不断完善国内分级连锁体系。目前公司眼科医院数量已达230余家，其中体内门店约80家。2017年门诊量508万人次，同比增长36.99%；手术量517，613例，同比增长37.21%。后续随着消费升级和门店下沉，公司服务人次有望持续提升。

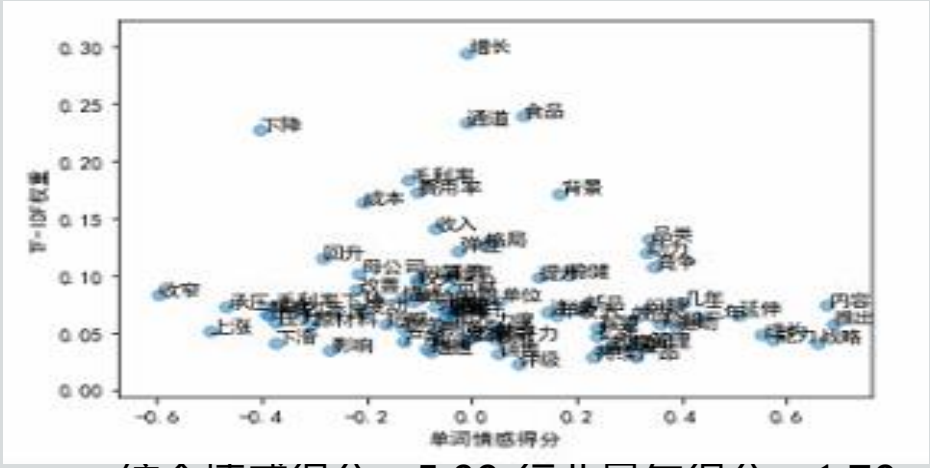
员工激励充分，品牌力持续提升。公司多层次激励体系日趋完善，通过限制性股票激励、“省会医院合伙人计划”等充分调动员工积极性。公司学术体系完善，并不断构建全球化的眼科平台，品牌力持续提升。

盈利预测：预计公司2018-2020年EPS分别为0.62元、0.81元、1.04：元，对应PE分别为68X、52X、41X，首次覆盖给予“买入”评级。<!- 爱尔眼科（300015.SZ）：业绩符合预期，高增长有望持续- >null

研报分类模型——案例分析

竞争格局改善+头部公司





综合情感得分：5.09 行业景气得分：1.70
份额提升得分：1.70 产品涨价得分：1.70

事件描述：安井食品披露2017年年报，主要内容如下：2017年实现收入34.84亿元，同比增长16.27%，归属于母公司净利润2.02亿元，同比增长14.11%，其中2017Q4实现收入10.46亿元，同比增长15.58%，归属于母公司净利润6332.97万元，同比增长26.51%。

事件评论：收入稳健增长，吨价进入提升通道：2017年收入同比增长16.27%，分产品看，速冻鱼糜制品/速冻肉制品/速冻其他食品/速冻面米制品分别同比增长16.71%/9.43%/27.34%/19.68%，速冻面米制品增速较突出，2017年新品小龙虾贡献收入300多万。从量价看，速冻鱼糜制品量增长16.2%，价增长0.41%，速冻肉制品量增长5.3%，价增长4%，速冻面米制品量增长13.6%，价增长5.4%，过去几年公司为了**抢占市场份额**以及成本下降的背景下，促销力度较大，吨价自2014年开始，连续下降三年，2017年公司吨价进入回升通道。

成本压力下毛利率短期承压：2017年毛利率同比下降0.8pct至26.3%，净利率同比下降0.1pct至5.8%，在**单价提升**背景下，毛利率下降，主要是由于原材料成本上涨，且公司处于产能扩张阶段，单位制造费用亦有所上升，影响毛利率。分季度看2017Q1/Q2/Q3/Q4毛利率分别同比变动-2.1pct/0.5pct/-1.5pct/-0.3pct，预计四季度毛利率同比下滑幅度收窄，主要是由于吨价提升贡献。2017年公司期间费用率同比下降0.7pct，其中销售费用率/管理费用率/财务费用率分别同比下降0.1pct/0.5pct/0.1pct，主要是由于**规模效应**。

涉足小龙虾再添增长动力，**竞争格局改善**背景下，盈利能力弹性大：2017年推出调味小龙虾，公司过去主营为火锅料，通过面点证明了其品类延伸能力，小龙虾有望成为第三大战略品类，中期"火锅料+面点+小龙虾"三驾马车驱动，成长动力再添筹码。同时认为随着成本上行，竞争格局有望持续改善，公司吨价提升空间大（2013-2016年吨价下降22.5%），毛利率和净利率均有较大弹性。预计2018/2019年EPS分别为1.16/1.46元，2018年PE仅22倍，维持"买入"评级。<!- 安井食品（603345.SH）：收入利润稳健增长，吨价进入回升通道- >

研报分类模型--监督学习

标注数据集

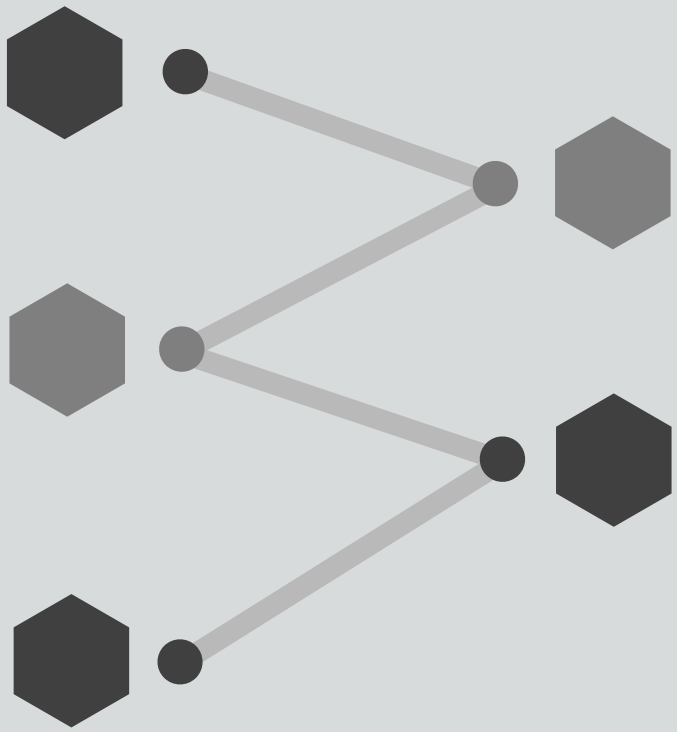
通过关键词检索与人工标注的方法标注部分数据集作为机器学习训练集。

划分数据集

将数据集划分为训练集与测试集。

预测数据

使用训练好的模型对预测集进行预测。



TF-IDF提取文档特征

基于TF-IDF方法，构造文档特征

训练模型

运用主流监督学习方法，训练文档分类模型。

FOUR

改进与提升

改进与提升

优化数据预处理

继续补充停用词及自定义词典，避免一些专业术语被错误划分导致后续情感得分出现偏误。

补充情感种子词典

基于文献中给出的CFSD词典，我们可以使用研报的常用情感表达词汇，更新种子词典。

优化情感得分

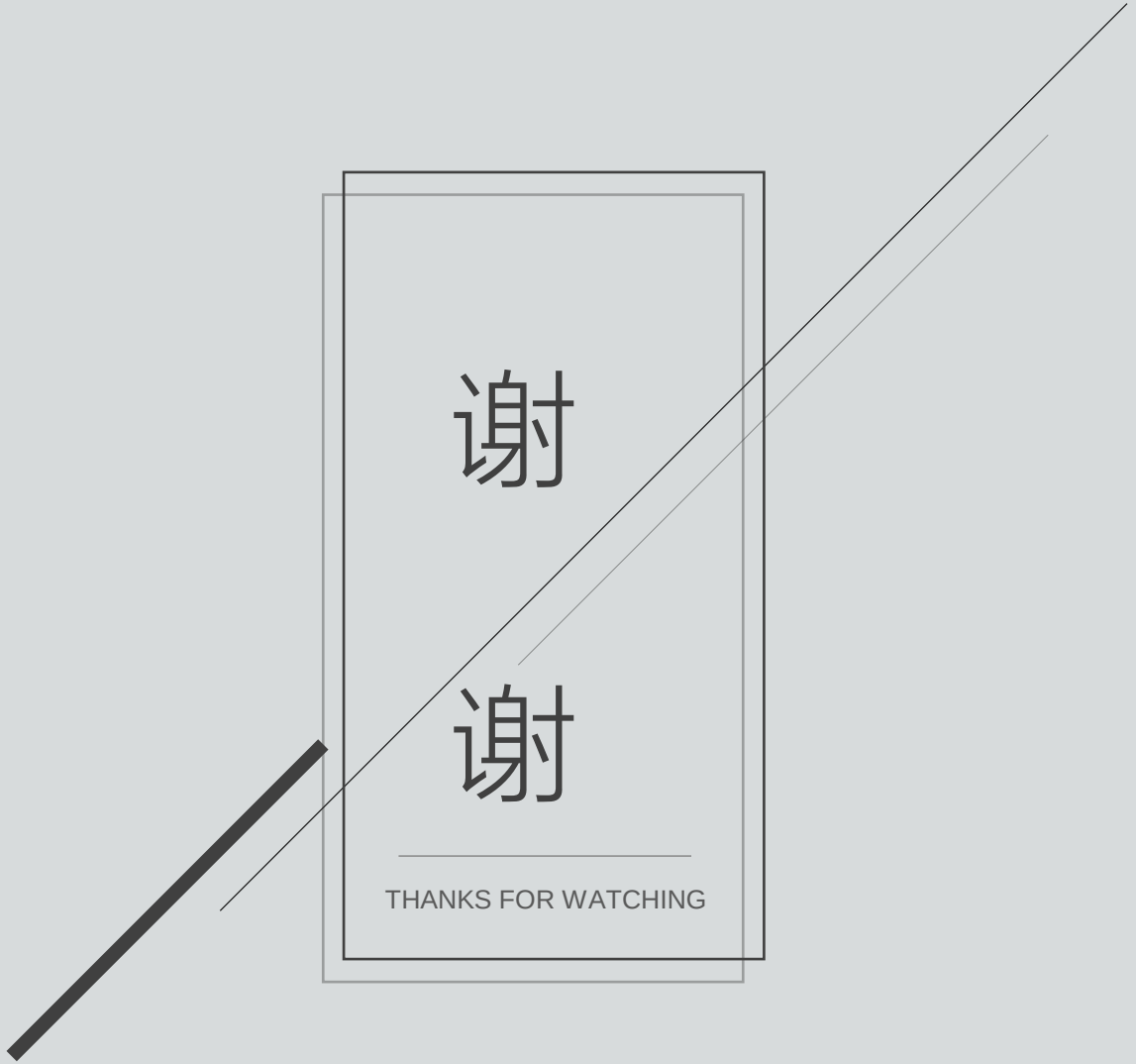
改进SO-PMI算法或者使用非对称的S型函数进行映射（归一化），对消极词汇进行惩罚，改善情感得分分布明显右偏的状况。

补充关键词索引

针对文本分类的5个类别，后续我们可以继续补充关键词，提高分类的准确性。

优化分类模型效果

模型仅使用机器学习框架训练模型，后续可以考虑使用深度学习框架训练模型。


谢
谢
THANKS FOR WATCHING